C | A | U Christian-Albrechts-Universität zu Kiel Technische Fakultät

CDS 303: Scientific Data Mining

Winter Term 2018/19

Lecture 2: Data preprocessing and feature spaces

Lectures: Prof. Dr. Matthias Renz

Exercises: TBA

Recap from previous lecture

- KDD definition
- KDD process
- DM step
- Supervised (or predictive) vs Unsupervised (or descriptive) learning
- Main DM tasks
  - Clustering: partitioning in groups of similar objects
  - Classification: predict class attribute from input attributes, class is categorical
  - Regression: predict class attribute from input attributes, class is continuous
  - Association rules mining: find associations between attributes
  - Outlier detection: identify non-typical data

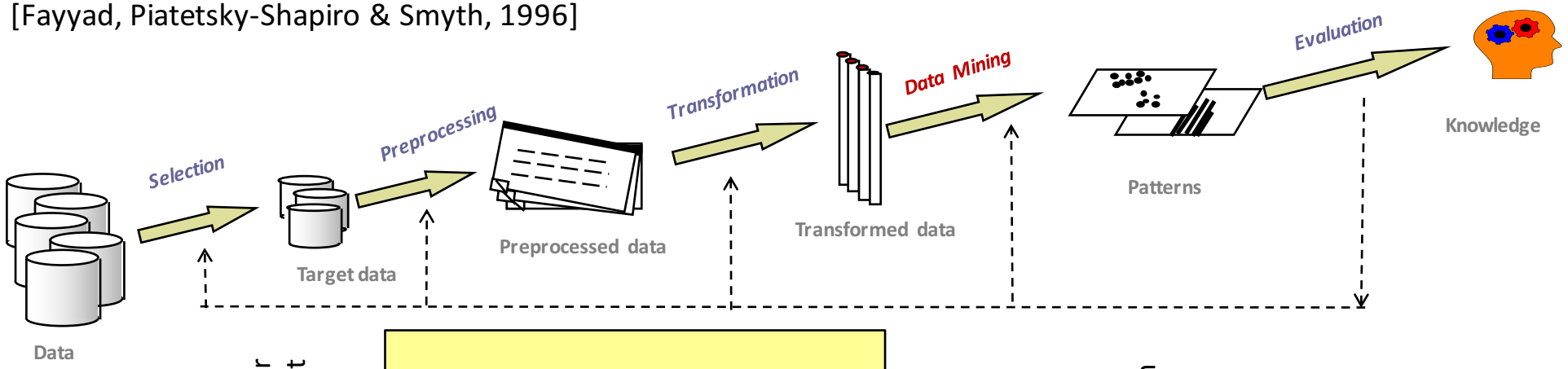
How data mining differs from database querying?

Outline

- Data preprocessing
- Decomposing a dataset: instances and features
- Basic data descriptors
- Feature spaces and proximity (similarity, distance) measures
- Feature transformation for text data
- Homework/ Tutorial
- Things you should know from this lecture

Recap: The KDD process

[Fayyad, Piatetsky-Shapiro & Smyth, 1996]



Selection:
- Select a relevant dataset or focus on a subset of a dataset
- File / DB/

Preprocessing/Cleaning:
- Integration of data from different data sources
- Noise removal
- Missing values

Transformation:
- Select useful features
- Feature transformation/ discretization
- Dimensionality reduction

Data Mining:
- Search for patterns of interest

Evaluation:
- Evaluate patterns based on interestingness measures
- Statistical validation of the Models
- Visualization
- Descriptive Statistics

Why data preprocessing?

- Real world data are noisy, incomplete and inconsistent:
  - Noisy: errors/ outliers
    - erroneous values : e.g. salary = -10K
    - unexpected values: e.g. salary=100K when the rest dataset lies in [30K-50K]
  - Incomplete: missing data
    - missing values: e.g., occupation=“ ”
    - missing attributes of interest: e.g. no information on occupation
  - Inconsistent: discrepancies in the data
    - e.g. student grade ranges between different universities might differ, in DE [1-5], in GR [0-10]
- “Dirty” data → poor mining results
- Data preprocessing is necessary for improving the quality of the mining results !
- Not a focus of this class!

Know your data!

Major tasks in data preprocessing

- Data cleaning:
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data integration:
  - Integration of multiple databases, data cubes, or files (Entity identification, Value resolution)
- Data transformation:
  - Normalization in a given range, e.g., [0-1]
  - Generalization through some concept hierarchy, e.g. "*milk 1.5% brand x*"→"*milk 1.5%*" or "*milk*"
- Data reduction:
  - Aggregation, e.g., from 12 monthly salaries to month's average salary.
  - Dimensionality reduction, through e.g., PCA
  - Duplicate elimination

Outline

- Data preprocessing
- Decomposing a dataset: instances and features
- Basic data descriptors
- Feature spaces and proximity (similarity, distance) measures
- Feature transformation for text data
- Homework/ Tutorial
- Things you should know from this lecture

Datasets = instances + features

- Datasets consists of instances (also known as examples or objects)
  - e.g., in a university database: students, professors, courses, grades,…
  - e.g., in a library database: books, users, loans, publishers, ….
  - e.g., in a movie database: movies, actors, director,…
- Instances are described through features (also known as attributes or variables)
  - E.g. a course is described in terms of a title, description, lecturer, teaching frequency etc.
  - An easy to visualize example: if our data are in a database table, the rows are the instances and the columns are the features.

ID	Gender	Height(cm)	Weight (kg)	Hair Color	Blood Group	Glasses	Smoker	GGS 787 Grade
67	Female	175	60	brown	A	no	frequent	A+
68	Female	176	52	blond	AB	yes	frequent	A
69	Female	176	63	black	A	yes	casual	A+
70	Female	179	65	brown	0	yes	no	B

Basic feature types

- Binary/ Dichotomous variables
- Categorical (qualitative)
  - Nominal variables
  - Ordinal variables
- Numeric variables (quantitative)
  - Interval-scale variables
  - Ratio-scaled variables

Binary/ Dichotomous variables

- The attribute can take two values, {0,1} or {true,false}
  - usually, 0 means absence, 1 means presence
  - e.g., smoker variable: 1→ smoker, 0→ non-smoker
  - e.g., true (1), false (0)
- Symmetric binary: both outcomes equally important:
  - e.g., gender (male, female)
- Asymmetric binary: outcomes not equally important.
  - e.g., medical tests (positive vs. negative)
  - Convention: assign 1 to most important outcome (e.g., HIV positive)

Person	isSmoker
Eirini	0
Erich	1
Kostas	0
Jane	0
Emily	1
Markus	0

Give me some examples of binary variables!

ID	Gender	Height(cm)	Weight (kg)	Hair Color	Blood Group	Glasses	Smoker	GGS 787 Grade
67	Female	175	60	brown	A	no	frequent	A+
68	Female	176	52	blond	AB	yes	frequent	A
69	Female	176	63	black	A	yes	casual	A+
70	Female	179	65	brown	0	yes	no	B

Categorical: Nominal variables

- The attribute can take values within a set of *M* categories/ states.
  - *No ordering* in the categories/ states.
  - Only *distinctness relationships*, i.e., *equal (==)* and *different* (!=), apply.
  - Examples:
    - Colors = {brown, green, blue,…,gray},
    - Occupation = {engineer, doctor, teacher, …, driver}
    - Gender = {male, female]

Person	gender	occupation
Eirini	female	professor
Erich	male	engineer
Kostas	male	doctor
Jane	female	engineer
Emily	female	teacher
Markus	male	driver

Give me some examples of nominal variables!

ID	Gender	Height(cm)	Weight (kg)	Hair Color	Blood Group	Glasses	Smoker	GGS 787 Grade
67	Female	175	60	brown	A	no	frequent	A+
68	Female	176	52	blond	AB	yes	frequent	A
69	Female	176	63	black	A	yes	casual	A+
70	Female	179	65	brown	0	yes	no	B

Categorical: Ordinal variables

- Similar to categorical variables, but the *M* states are ordered/ ranked in a meaningful way.
  - There is an *ordering* between the values.
  - Allows to apply *order relationships*, i.e., $>, \geq, <, \leq$
  - However, the difference and ratio between these values has no (quantitative) meaning.
  - Examples:
    - School grades: *{A,B,C,D,F}*
    - Movie ratings: {hate, dislike, indifferent, like, love}
      - Also, movie ratings: {*, **, ***, ****, *****}
      - Also, movie ratings: {1, 2, 3, 4, 5}
    - Medals = {bronze, silver, gold}

Person	A beautiful mind	Titanic
Eirini	5	3
Erich	5	1
Kostas	3	3
Jane	1	5
Emily	1	5
Markus	4	3

Give me some examples of ordinal variables!

ID	Gender	Height(cm)	Weight (kg)	Hair Color	Blood Group	Glasses	Smoker	GGS 787 Grade
67	Female	175	60	brown	A	no	frequent	A+
68	Female	176	52	blond	AB	yes	frequent	A
69	Female	176	63	black	A	yes	casual	A+
70	Female	179	65	brown	0	yes	no	B

Numeric: Interval-scale variables

- Measured on a scale of equal-sized units
  - It is assumed that the intervals keep the same importance throughout the scale.
- Differences between values are meaningful
  - The difference between 90° and 100° temperature is the same as the difference between 40° and 50° temperature.
- Ratio still has no meaning
  - A temperature of 2° Celsius is not much different than a temperature of 1° Celsius.
  - The issue is that the 0° point of the Celsius scale is in a physical sense arbitrary and therefore the ratio of two Celsius temperatures is not physically meaningful.
- No meaningful (unique and non-arbitrary) zero value
- Examples:
  - Temperature in Farenheit or Celsius
  - Calendar dates

Give me some examples of interval-scale variables!

Numeric: Ratio-scale variables

- Both differences and ratios have a meaning
  - E.g., a 100 Kgs person is twice heavy as a 50 Kgs person.
  - E.g., a 50 years old person is twice old as a 25 years old person.
- Meaningful (unique and non-arbitrary) zero value
- Examples:
  - age, weight, length, number of sales
  - temperature in Kelvin
    - When measured on the Kelvin scale, a temperature of 2^o is, in a physical meaningful way, twice that of a 1^o.

Give me some examples of ratio-scale variables!

ID	Gender	Height(cm)	Weight (kg)	Hair Color	Blood Group	Glasses	Smoker	GGS 787 Grade
67	Female	175	60	brown	A	no	frequent	A+
68	Female	176	52	blond	AB	yes	frequent	A
69	Female	176	63	black	A	yes	casual	A+
70	Female	179	65	brown	0	yes	no	B